Resource Construction and Ensemble Learning Based Sentiment Analysis for the Low-resource Language Uyghur

Azragul Yusup[1,2], Degang Chen[1], Yifei Ge[1], Hongliang Mao[1], Nujian Wang[1]*

[1] College of Computer Science and Technology, Xinjiang Normal University, China
[2] National Language Resource Monitoring & Research Center of Minority Languages, China
Azragul2010@126.com, iloveqq001@vip.qq.com, 1453259830@qq.com, Wrinkle_625@163.com, lxznjw@163.com

Abstract

To address the problem of scarce low-resource sentiment analysis corpus nowadays, this paper proposes a sentence-level sentiment analysis resource conversion method HTL based on the syntactic-semantic knowledge of the low-resource language Uyghur to convert high-resource corpus to low-resource corpus. In the conversion process, a k-fold cross-filtering method is proposed to reduce the distortion of data samples, which is used to select high-quality samples for conversion; finally, the Uyghur sentiment analysis dataset USD is constructed; the Baseline of this dataset is verified under the LSTM model, and the accuracy and F1 values reach 81.07% and 81.13%, respectively, which can provide a reference for the construction of low-resource language corpus nowadays. The accuracy and F1 values reached 81.07% and 81.13%, respectively, which can provide a reference for the construction of today's low-resource corpus. Meanwhile, this paper also proposes a sentiment analysis model based on logistic regression ensemble learning, SA-LREL, which combines the advantages of several lightweight network models such as TextCNN, RNN, and RCNN as the base model, and the meta-model is constructed using logistic regression functions for ensemble, and the accuracy and F1 values reach 82.17% and 81.86% respectively in the test set, and the experimental results show that the method can effectively improve the performance of Uyghur sentiment analysis task.

Keywords: Low-resource language, Uyghur, HTL, Stacking ensemble learning, Sentiment analysis

1 Introduction

With the continuous development of artificial intelligence, artificial intelligence systems with natural language understanding have brought great convenience to people. Most of today's natural language understanding systems rely on a large number of language annotations and only provide applications for high-resource languages; they do not have good generalization performance for small languages with low resources and scarce language annotations. Uyghur [1] falls squarely into the low-resource language category and does not have sufficient resources to exploit. Sentiment analysis [2], as one of the important branches of natural language processing, is the main research task to discern the sentiment of a user's opinion based on the semantic information of the text. Combining the above two points, this study focuses on building a low-resource Uyghur sentiment analysis corpus, and using deep learning techniques and Stacking ensemble learning techniques to build a sentiment analysis model SA-LREL to realise the sentiment classification task, which can provide a theoretical basis for subsequent natural language processing tasks related to low-resource languages.

The main contributions of this paper are as follows.

(1) In response to the present-day shortage of low-resource corpus, this paper proposes a HTL method for converting high-resource corpus data to low-resource sentiment corpus. This method can generate low-resource high-quality sentence-level data samples with high efficiency, which can provide a reference for low-resource sentence-level corpus construction.

(2) In the HTL method, resource data conversion is required, and in this process, a certain amount of data distortion may be caused. For this reason, this paper proposes a k-fold cross data filtering method, which refers to the machine learning idea and has a certain theoretical basis, and can better filter out the data samples with higher quality, thus reducing the accuracy loss of the conversion process.

(3) In the study of sentiment analysis, this paper proposes a Sentiment analysis model based on logistic regression ensemble learning (SA-LREL). The model was experimentally evaluated on a constructed low-resource Uyghur dataset USD. The base models selected for the ensemble model are all lightweight network models, such as TextCNN, LSTM, RCNN, etc., and the meta-model are chosen as logistic regression functions. This can improve the model performance in the case of sparse low-resource data samples and does not over-fit due to the small amount of data. This model is used for sentiment classification of linguistic Uyghur monolingual, which can provide a theoretical study for low-resource sentiment analysis.

2 Related Work

Low-resource language comprehension tasks have become an important research hotspot nowadays, and the sparse corpus related to low-resource languages does not

*Corresponding Author: Azragul Yusup; E-mail: Azragul2010@126.com
DOI: 10.53106/160792642023072404018

meet the needs of language comprehension tasks for low-resource languages today. Different researchers have used different methods to obtain the resources of low resource language corpus. Dehkharghani et al. [3] based on WordNet's translation method, translated the existing polarity words in resource rich language (English) into resource poor language (Persian), and then used supervised learning method to estimate the polarity score of the translated Persian words, and finally constructed a tri-categorized polarity score Persian lexicon SentiFars. Hasmot et al. [4] used Microsoft Excel 2016 tool to transform the pre-processed high resource English data into a Bengali lexical corpus with the help of its GOOGLE TRANSLATE engine. This process required greater human proofreading to improve the accuracy of the transformed data and the final corpus form was aligned with the SentiWordNet dataset for subsequent researchers.

Deep learning techniques have won a new life in recent years, and many scholars in today's research have realized its superiority and focused on applying deep learning to low resource dimensional language understanding tasks. Wang Shuheng et al. [5] used the skip-gram model in word2vec to convert Uyghur words into low-dimensional word vectors that can be understood by computers. Meanwhile, the input network incorporates sentiment word features for training, and BiLSTM bidirectional network learning is selected for feature extraction. It is experimentally verified to be more effective compared to traditional CNN, RNN, SVM and other models. Yimamu Aishan et al. [6] implemented five sentiment classifications based on deep belief networks for sentence-level text, such as happiness, anger, sadness, joy, and objectivity, and the sentiment features were extracted from eight sentiment features such as lexicality, emotion words, degree adverbs, and negation words, and the multiple features were fused and trained by DBN (Deep Boltzmann machine) model. The results show that the DBN network with three layers of RBM superposition achieves the best classification results. The paper also corroborates that deep learning models extracting shallower feature representations are more suitable for text sentiment classification tasks. Yang et al. [7] proposed a multi-headed attention-based capsule network model for personal pronoun analysis in Uyghur language, using a multi-headed attention mechanism to extract multi-level semantic features of personal pronouns and candidate antecedents, and an IndRNN network on the other side to extract sentence semantic features. The paper The constructed capsule network model achieves better results in the application of personal pronoun analysis for low-resource languages. mBERT model [8] although achieves better results in cross-language transfer learning, this is mostly based on the first 104 languages pre-trained in the Wikipedia corpus. Although the model slightly oversamples its pre-trained data for low-resource languages, high-resource languages are still more prevalent, and therefore, there is still a problem leading to language imbalance, and the performance for zero-sample learning for 104 languages not included is not considered excellent, which indicates that there is still a need to increase the data resources related to low-resource languages. daniel et al. [9] and Dou et al. [10] introduced active learning and meta-learning algorithms for the comprehension task on low-resource languages, and experiments show that this class of methods is also significantly better than benchmark models such as BERT.

In summary, today's Uyghur sentiment corpus is scarce, and it is difficult to obtain a large number of sentence-level comments with sentiment polarity tags. Therefore, this paper proposes a resource-based transformation HTL method to transform the open-source high-resource sentiment corpus into a low-resource Uyghur sentiment corpus. Since the sentiment labels before conversion already have sentiment labels, the method can greatly reduce the cost of manual labeling; moreover, this paper proposes a k-fold cross-data filtering method that can assist in filtering out high-quality samples before data conversion and reduce the distortion caused by data conversion. After data transformation combined with a simple review by Uyghur language experts, a final high-quality low-resource language-Uyghur sentiment analysis dataset USD can be constructed. Meanwhile, this paper proposes a logistic regression-based ensemble learning model for Uyghur sentiment analysis, which is experimented on the constructed Uyghur sentiment analysis dataset USD to verify the effectiveness of this model.

3 Uyghur Sentence-level Sentiment Analysis Dataset Construction

Due to the scarcity of resources in today's Uyghur corpus, sentence-level data samples that can be applied to user sentiment analysis tasks are even scarcer. Even if large-scale crawling is used, it still requires a lot of effort for feature engineering, e.g., manual cleaning of the dataset. Moreover, for the construction of low resource dimensional sentiment analysis corpus, it is also necessary to have certain linguistic knowledge to be able to screen data with sentiment polarity viewpoints; and then to manually label the data to obtain the final sentiment analysis dataset, and the financial resources spent on the above works cannot be underestimated. Therefore, this paper proposes a sentiment analysis corpus construction method based on mapping high resource language (Chinese) to low resource language (Uyghur).

3.1 High Resource to Low Resource Conversion Method (HTL Method)

The HTL resource construction method is shown in the following Figure 1:

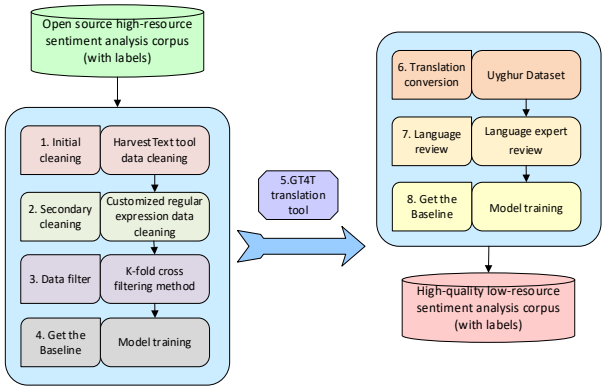


Figure 1. HTL sentence-level resource construction method

1. **Initial cleaning**: In the preliminary stage, a suitable open high resource sentence-level sentiment analysis corpus needs to be selected for transformation. After that, the data is then cleaned using the open source tool HarvestText, which is a text processing analysis library focusing on weak supervision. It can be used for named entity recognition, fine-grained clause splitting, dependent syntactic analysis, text cleaning and many other aspects, and this study only applies this function of its text cleaning

2. **Secondary cleaning**: In the initial cleaning of the data, a weakly supervised approach is used for processing throughout. Therefore, in the secondary cleaning, we check the data by randomly selecting samples of the data and use custom regular expressions to further clean up the dirty data that have not yet been identified in the data.

3. **Data filters**: The data filters use the K-fold cross-filtering method proposed in this paper, using the idea of non-repetitive sampling, to select high-quality corpus with clear semantic expressions of emotion, which is convenient for subsequent resource transformation.

4. **Get the Baseline**: The LSTM and BERT network models are selected as the baseline models in this method, and the cleaned high-resource corpus dataset is trained and tested to obtain the Baseline of this data.

5. **GT4T tool**: In the data conversion study, the translation software used in this HTL method is the GT4T tool, which supports 375 languages and has good adaptability for Uyghur.

6. **Translation conversion**: The high-resource data was organized into Excel file format, and the filtered high-resource corpus data were transformed into low-resource corpus Uyghur data by GT4T tool.

7. **Language review**: None of the preliminary work in this method requires the assistance of a low-resource language expert. The use of unsupervised translation in the previous step has already reduced the amount of language translation work. This stage requires only a simple check by the linguist to correct a small number of semantic expression errors and amend the error labels, and a good return is achieved.

8. **Get the Baseline**: In the final stage, a suitable language model is selected and the processed low-resource corpus is trained to get the Baseline of the constructed corpus, so that the construction of the low-resource language Uyghur sentiment analysis corpus is completed.

3.2 K-fold Cross-filtering Method

This study proposes the k-fold cross-filtering method with reference to the commonly used k-fold cross-validation method in machine learning. The role of using k-fold cross-validation method is mostly used to adjust the model parameters under the condition of less samples, so as to obtain the optimal model. In contrast, the purpose of the k-fold cross-filtering method proposed in this study is to screen high-quality high-resource corpus data with clear semantic expressions that are more easily distinguishable. This helps to reduce the loss of the overall corpus benchmark caused by data distortion during the high-resource to low-resource transformation phase. k-fold cross-filtering method steps are shown in the following Figure 2.

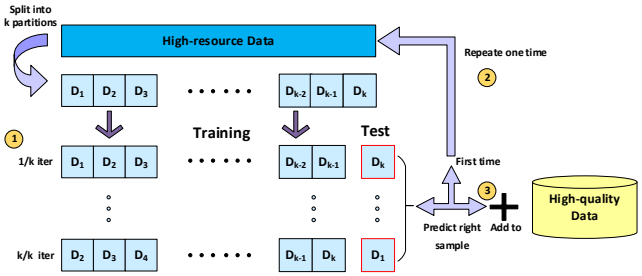


Figure 2. k-fold cross-filtering method

As shown in the Figure 2, in the first step, the dataset is divided by k-fold, k-1 data are used for model training and 1 data is used for testing. In this study, the data with correct predictions in the test dataset are screened out, and in the second step, the dataset continues to be cross-separated by k-fold and sent to the model training, and in the third step, the data with correct predictions from the test set are screened out again and added to the corpus. It is worth mentioning that in order to prevent the phenomenon of model overfitting due to repeated training, in the process of repeated training, this paper chooses to train from scratch and does not load the last training weights.

4 Advantages of Ensemble Learning

Ensemble learning algorithm is a technique that can effectively improve the accuracy of models in various natural language understanding tasks. It learns by integrating multiple base classifiers for a specific downstream task. Base classifiers usually use weakly supervised classifiers, and the accuracy of model prediction can be greatly improved by using the ensemble learning approach. Due to the weakly supervised learning approach of a single base classifier itself, it limits the upper limit capability of its recognition, and its recognition correct rate is higher than that of a random guess polynomial algorithm, but it is still slightly inadequate to be applied to the learning of a specific downstream task. Therefore, the idea of ensemble learning was born; it makes the ensemble model more robust by integrating multiple weakly supervised models; it is expected to correct the few weakly supervised classifiers with incorrect prediction results in the ensemble model by using the voting idea of “majority voting fusion”.

5 Sentiment Analysis Model Based on Logistic Regression Ensemble Learning, SA-LREL

The ensemble model construction in this paper uses an ensemble framework with a hierarchical structure, Stacking [11]. The first layer input of the ensemble model requires the original dataset for training and is composed using multiple base learners. The second layer integrates the prediction results of all models in the upper layer as the input of this layer, which is used for training in the current layer. In this

paper, the main research focus is on the sentiment analysis task of low resource language Uyghur with less data volume, so the base models are all selected for ensemble with small number of parameters, which can effectively way overfitting. The structure of SA-LREL, a sentiment analysis model based on logistic regression ensemble learning, is shown in Figure 3.

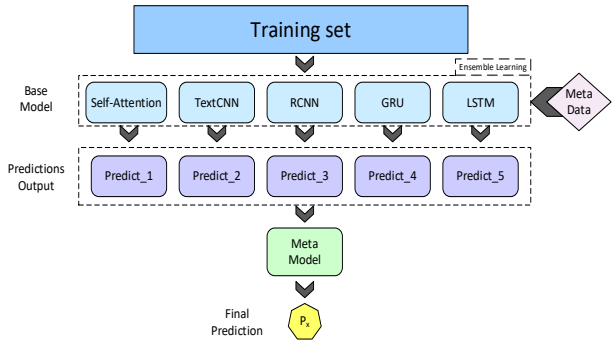


Figure 3. Sentiment analysis model based on logistic regression ensemble learning, SA-LREL

The base models in the ensemble model proposed in this paper were selected as the feature extractors of the original data from lightweight convolutional networks, self-attention mechanism [12] networks, recurrent neural networks, and RCNN [13] net-works, respectively. The reasons for choosing the above base models are as follows.

1. The convolutional network TextCNN [14] uses a sliding window form for convolution, and different sizes of convolution were used for learning in [3-5], which has better ex-traction ability for local features, obtaining n-gram language information, and high model learning efficiency.

2. The self-attention mechanism network is able to tend the focus of feature learning to local information, while ignoring irrelevant information. For example, the serialized input samples are assigned different weights to focus on the features related to emotional information. This network is selected in the base model to be able to learn the important emotional features in the sequence.

3. And in recurrent neural networks, two models, LSTM and GRU, are selected in this paper as feature learners for the original data, each with its own advantages and disadvantages; both are variants of RNN, and the corresponding improvements are made RNN to solve the problem that traditional recurrent neural networks may have semantic loss and training gradient disappearance for longer information; LSTM model consists of forgetting gate, input gate, output gate The LSTM model consists of forgetting gates, input gates, and output gates to control and protect the activation state of neurons; GRU is improved on the basis of LSTM and only retains reset gates and update gates. Compared with GRU, LSTM requires more training parameters and has stronger performance. However, the GRU model has fewer training parameters and also accelerates the shortening of the training cycle, which is more simplified and can converge faster to achieve the desired effect of the model. Therefore, in this paper, both are added to the ranks of base models for ensemble.

4. The RCNN network is built by combining the advantages of CNN and RNN, first using CNN convolutional structure as the front-end feature extraction network, and then adding the bi-directional RNN network Bi-LSTM to extract the contextual information to obtain deeper semantic information.

After that, each base model is trained to output the prediction results as Predict_1~ Predict_5, and all the prediction results are ensemble as the feature input of the meta-model.

Finally, on the meta model, logistic regression functions are selected in this paper for learning the features trained by the base models to achieve the task of sentiment analysis for the low-resource language Uyghur.

6 Experimental Analysis & Evaluation

6.1 Experimental Evaluation Metrics

In this paper, we focus on the construction of a low-resource sentiment analysis corpus and the construction of a sentiment analysis model. To test the quality performance of the low-resource corpus construction and the effectiveness of the sentiment analysis model SA-LREL classification, we selectively used Accuracy and F1-Measure metrics for evaluation. The main components of the metrics measuring the classification model can be represented by the confusion matrix shown in Table 1, which indicates the correlation between the classification of the true labeling situation of the sample set and the actual predicted outcome classification.

Table 1. Confusion matrix

Positive		Predicted	
		Negative	
Actual	Positive	True Positive (TP)	False Negative (FN)
	Negative	False Positive (FP)	True Negative (TN)

Where TP indicates the number of positive category samples that can be correctly predicted as positive by the model; similarly, TN indicates the number of negative category samples that can be correctly predicted as negative by the model; FP indicates the number of negative category samples that can be incorrectly predicted as positive by the model; and FN indicates the number of negative category samples that can be incorrectly predicted as negative by the model. The following four metrics are model measures derived from these four sample prediction results.

Accuracy is the proportion of correct predictions made by the model. It is calculated as shown below.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}. \quad (1)$$

Precision is expressed as the proportion of the number of samples predicted by the model to be positive that are actually positive. It is calculated as shown below.

$$\text{Precision} = \frac{TP}{TP + FP}. \quad (2)$$

Recall is expressed as the number of all positive samples in the sample set and the proportion of samples that are actually predicted to be positive. The formula for its calculation is shown below.

$$\text{Recall} = \frac{TP}{TP + FN}. \quad (3)$$

In real-life experiments, both Precision and Recall values can be too high or too low for the model to perform well. Therefore, the F1 value is quoted on the basis of the two. Algorithmically, it is a weighted average of the two, which effectively solves the difficulty of evaluation caused by the unbalanced distribution of the sample set, etc. The F1 value is a comprehensive evaluation of the Precision and Recall values, and the β coefficient set in this experiment is 1, and its calculation formula is shown as follows.

$$F_\beta = \frac{(1+\beta^2)\cdot \text{Precision} \times \text{Recall}}{(\beta^2 \cdot \text{Precision}) + \text{Recall}}. \quad (4)$$

6.2 Experimental Platform and Parameter Settings

The operating system used in the experimental environment of this paper is Ubuntu 20.04, the graphics card is RTX3080, the memory is 32 GB, the CUDA version is 10.2, and the Tensorflow version is 1.14.0. The data cleaning process of the high-resource part of the model uses the HarveText tool to process the constructed sentiment analysis data throughout the experiment, and the training settings are The maximum length of the input samples was set to 128, the size of Batch_Size in training was 32, the Dropout parameter was 0.7, and the ratio of random data division was 8:1:1.

6.3 Corpus Resource Assessment

6.3.1 High Resource Corpus Evaluation

Since the low resource target language of this study is Uyghur, Uyghur is a common language of expression within China. Uyghur has a strong correlation with the content of Chinese language expressions. Therefore, it is most appropriate to select a Chinese high-resource corpus. The high-resource dataset selected for this study is the open-source microblog dataset weibo_senti_100k, and HarvestText is used for initial cleaning of the data. The tool refers to many tools such as snownlp, pyhanLP, funNLP, etc. It provides the function of Chinese text cleaning, which can effectively clean up "@", emoji, URL, email, html code and other special characters in microblogs, as well as traditional to simplified characters. The benchmark models for the sentiment analysis task on Chinese data were chosen to evaluate LSTM and BERT, and the data division ratio was 8:1:1. The comparison of before and after data cleaning is shown in the following Table 2.

Table 2. Comparison of the accuracy of test sets before and after cleaning of high resource corpus data

No.	Data set type	Chinese (Accuracy%)	
		LSTM	BERT
1	Original data	95.48	97.81
2	Cleaned data	74.68	77.80
3	Data baseline after K-fold cross-filtering	92.33	92.56

As shown in the table above, the raw, unwashed Chinese data in No. 1 were trained on the LSTM and BERT models, respectively, and although the accuracy in the test set reached 95.48% and 97.81%, respectively, the reason for this illusion is the following problem with this data: the high-resource sentiment analysis dataset used in this study was constructed based on the content of microblogs, and there are many microblog-specific emoticons in the text feature symbols; therefore, some of the features learned by the model are automatically classified based on the emoji in the text, such as the emoji "[xi xi]", which represents positive sentiment; so the model can easily learn these features. Even because it only classifies based on emojis and completely disregards the semantic information originally expressed by the text, this can be highly biased. And the classifier learned by the model will not work for the content of microblog texts with contradictory views and emojis, or posts without emojis.

In No. 2, this study processed the textual information of this data appropriately: de-weighting, de-dirtying, removing relevant emoticons and meaningless special characters, etc. At this point, the accuracy of both models was reduced, 74.68%,77.8% respectively. However, at this time, the model really learns the semantic information expressed by the text, which can effectively improve the model robustness. The model has better generalization to new samples that the model has not been exposed to.

In No. 3, this study further processed the high-resource data by manually sampling and modifying the data mislabeling and removing the data with unclear sentiment expression. It is beneficial to improve the semantic expression error of the subsequent conversion of the high-resource language (Chinese) sentiment analysis data to high-resource language (Uyghur) sentiment analysis data. About eight thousand data were randomly selected for resource conversion.

6.3.2 Validity of K-fold Cross-filtering Method

In this section, the high resource sentiment analysis corpus after secondary processing in the previous section was selected for experiments to verify the effectiveness of the k-fold filtering method proposed in this study. The following Table 3 compares the effects of the two data filtering methods on the benchmark of the high-resource corpus.

Table 3. Comparison of the accuracy of test sets with different data filtering methods

No.	Data filtering method	Chinese (Accuracy%)	
		LSTM	BERT
3.1	Traditional random loop filtering method	89.57	90.71
3.2	K-fold cross-filtering method	92.33	92.56

No. 3.1 used the conventional random data filtering method, which randomly disrupted the data set after dividing the data according to the ratio of 8:1:1; iterated on the LSTM model for 50 rounds, and the number of Epochs for each round of model training was 3; the data with correct predictions in the test set were filtered out cyclically without keeping the model parameters. From the experimental evaluation on LSTM and BERT, it can be seen that this method has some randomness, and its accuracy rate is not theoretically justified although it has improved greatly after removing some of the dirty data and de-weighting, 89.57% and 90.71%, respectively.

No. 3.2 used the k-fold cross-filtering method proposed in this paper to divide the data by 5 folds, using the no-repeat sampling method, so that each sample is divided into the validation set or test set when training, and no duplication is possible. This cross-filtering method using sampling has a better theoretical basis, and the number of training rounds is small, there is no duplicate data, and the test set accuracy can reach 92.33% and 92.56% in LSTM and BERT, respectively, which is significantly higher than the previous filtering method.

6.3.3 Low Resource Corpus Evaluation

After conversion to low-resource language dimensional language data, since sufficient relevant pre-trained word vectors are not available, here the training model tests are evaluated with LSTM only. The experiments on the converted low-resource data are shown in the following Table 4.

Table 4. Comparison of test set accuracy after conversion of low resource corpus data

No.	Data set type	Uyghur (Accuracy %)
		LSTM
4	Converted low resource data	74.17
5	Baseline of the final data	81.07

No. 4, following the process in subsection 5 of this paper, was transformed by the GT4T tool to obtain the low resource dimensional language dataset. The accuracy is 74.17% when trained and validated on the LSTM model.

No. 5, the low-resource data were still filtered using K-fold cross-filtering, and the dataset was screened for training tests, while some samples with incorrect semantic expressions were modified by manual random sampling. Due to the experimental resource limitation, the final small sample Uyghur sentiment analysis dataset USD constructed in this study, it has 10,174 data items, including 5,076 negative samples and 5,098 positive samples. This dataset was trained on the LSTM model, and the accuracy rate was 81.07% on the test set, which can be used as a baseline for the construction of subsequent low-resource sentiment analysis datasets. The details of the sentiment analysis dataset for the low-resource corpus of Uyghur constructed in this study are shown in Table 5 below.

Table 5. Comparison of test set accuracy after conversion of low resource corpus data

Corpus	Data sets	Positive	Negative
Train	8139	4070	4069
Validation	1018	514	504
Test	1017	514	503
All	10174	5098	5076

6.4 Different Models Judged

In this paper, experiments are conducted on the construction of USD small sample Uyghur sentiment analysis dataset, and to ensure the persuasiveness and validity of the conclusions, the experiments are divided into datasets using the k-fold cross-validation method for experiments, and the final average of k sets of test results is taken as the estimation of model accuracy, where the value of K is 5. This experiment compares the accuracy of the predictions of each base model individually, and the final Stacking algorithm ensemble sentiment analysis model SA-LREL, the list of models is shown in Table 6 below.

No. 6, GRU, as one of the variants of recurrent neural network RNN, can solve the problem of long-term dependence of time series in traditional RNN, mainly consists of two gating structures, update gate and reset gate.

No. 7, LSTM, also as one of the variants of RNN, has more parameters and better performance compared to GRU, mainly consists of input gate, forgetting gate and output gate structures.

No. 8, self-Attention attention mechanism can effectively improve the model performance by focusing the feature extractor on the focused part of the input samples.

No. 9, TextCNN, Kim's proposed structure adopts three sizes of convolutional kernels, 3, 4 and 5, respectively, corresponding to n-tuple grammatical features, which have higher learning ability for short text-based shallow features, but do not perform well for longer text-based long-term dependency problems.

No. 10, RCNN, the structure is a combined CNN and RNN advantages of the hybrid neural net-work, the front-end first use CNN as feature extraction, and later use Bi-LSTM as far as possible to obtain the global information of the upper and lower layers, in order to improve the text classification performance as much as possible, more effective to make up for the structural shortcomings of RNN and CNN.

No. 11, SA-LREL is the model sentiment analysis model proposed by this paper, which combines the advantages of different network structures such as CNN and RNN to build, and the meta-model used for the subsequent sentiment classification task is a logistic regression function.

Table 6. Experimental comparison of different models

No.	Model	Uyghur	
		Accuracy (%)	F1 (%)
6	GRU	81.38	81.13
7	RCNN	81.01	81.31
8	Self-Attention	81.19	81.24
9	LSTM	81.07	81.13
10	TextCNN	81.48	81.44
11	Ours	82.17	81.86

Through the above experiments, the performance of the SA-LREL sentiment analysis model proposed in this paper can be verified, and the effectiveness of USD for Uyghur sentiment analysis dataset can be verified. Facing the current situation that the text lengths of the Uyghur dataset in low resource languages vary, the ensemble model we build has the following advantages in the sentiment analysis task: (1) it has the powerful feature extraction ability of CNN network and has better learning ability for some short texts; (2) it has the self-attentive mechanism network to focus on learning the sentiment features of long sequence texts; (3) it uses GRU, LSTM network to solve the long-sequence dependency problem and have better extraction ability for long texts; (4) RCNN combines the advantages of CNN and RNN for feature learning. Experiments show that the SA-LREL model proposed in this paper can achieve optimal results, and the accuracy and F1 values on the test set are 82.17% and 81.86%, respectively, which are significantly higher than other models, proving the effectiveness of this model in the field of sentiment analysis. The limited improvement in accuracy of this ensemble model may be due to the small number of data samples and the limitation of model feature learning capability. This study will be followed by further experimental validation by expanding more data samples and replacing the base model.

7 Conclusion

Faced with the problem of scarce low-resource sentiment analysis corpus, this paper proposes a novel sentence-level low-resource sentiment analysis dataset construction method, HTL, by converting the open-source high-resource Chinese sentiment analysis corpus into low-resource Uyghur data. To reduce the loss of text features due to sample distortion during data conversion, this paper proposes a k-fold cross-filtering method to select high-resource data samples with clearer semantic expressions for conversion, and finally constructs the low-resource Uyghur sentiment analysis dataset USD. After dividing the data by the 8:1:1 ratio, the accuracy and F1 values on the test set, using the LSTM model as the benchmark, are 81.07% and 81.13%. This method may provide a new solution for corpus construction in related low-resource languages.

In addition, this paper proposes a sentiment analysis model, SA-LREL, for the low-resource language Uyghur. The model uses an ensemble framework with Stacking hierarchical structure, and five lightweight networks, TextCNN, Self-Attention, RCNN, LSTM, and GRU, are used as the base model for ensemble training, and after the prediction output of the base model, the logistic regression function is used for secondary training to obtain the final sentiment prediction labels. Tested on the Uyghur dataset USD constructed in this paper, the accuracy and F1 values can reach 82.17% and 81.86% respectively, which can provide a theoretical basis for the study of low-resource sentiment analysis on small sample sets.

In the future, on the resource construction task, we will continue to improve the resource transformation HTL method, as well as extend it with more low-resource linguistic data samples for experimental studies. On the sentiment analysis task, we will further improve the generalisation capability of our models by fine-tuning their parameters, or using pre-trained models for ensemble, etc. We hope that our research will contribute to new developments in natural language processing for low-resource language understanding tasks.

Acknowledgment

This work was supported in part by Special Project for Construction of Innovation Environment (Talents and Bases) in the Autonomous Region - Natural Science Program (Special Cultivation of Ethnic Minority Science and Technology Talents) under Grant 2022D03001, in part by the Natural Science Foundation of China under Grant 61662081, and in part by the National Social Science Fund of China under Grant 14AZD11.

References

[1] L. Yan, *A Linguistic Ecological Study of China's Cross-border Languages' Planning in the Core Areas of the Belt and Road*, Ph. D. Thesis, Southwest University, Chongqing, China, 2018.

[2] D. M. E. D. M. Hussein, A survey on sentiment analysis challenges, *Journal of King Saud University-Engineering Sciences*, Vol. 30, No. 4, pp. 330-338, October, 2018.

[3] R. Dehkharghani, Sentifars: a Persian polarity lexicon for sentiment analysis, *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, Vol. 19, No. 2, pp. 1-12, March, 2020.

[4] H. Ali, M. F. Hossain, S. B. Shuvo, A. A. Marouf, BanglaSenti: A Dataset of Bangla Words for Sentiment Analysis, *2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, Kharagpur, India, 2020, pp. 1-4.

[5] S. Wang, T. Ibrahim, K. Abiderexiti, A. Wumaier, G. Abudouwaili, Sentiment classification of Uyghur text based on BLSTM, *Computer Engineering and Design*, Vol. 38, No. 10, pp. 2879-2886, October, 2017.

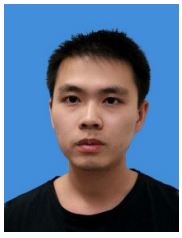
[6] Y. Abudoulikemu, M. Li, Z. Li, M. Chen, S. Tian, L. Yu, Emotional analysis of Uyghur sentence based on deep belief nets, *Application Research of Computers*, Vol. 35, No. 7, pp. 2066-2070, July, 2018.

[7] Q. Yang, L. Yu, S. Tian, J. Song, Multi-attention-based capsule network for Uyghur personal pronouns

resolution, *IEEE Access*, Vol. 8, pp. 76832-76840, April, 2020.
[8] Z. Wang, K. Karthikeyan, S. Mayhew, D. Roth, Extending multilingual BERT to low-resource languages, *Findings of the Association for Computational Linguistics: EMNLP*, Punta Cana, Dominican Republic (Online), 2020, pp. 2649-2656.
[9] D. Grießhaber, J. Maucher, N. T. Vu, Fine-tuning BERT for low-resource natural language understanding via active learning, *Proceedings of the 28th International Conference on Computational Linguistics (COLING)*, Barcelona, Spain (Online), 2020, pp. 1158-1171.
[10] Z. Y. Dou, K. Yu, A. Anastasopoulos, Investigating meta-learning algorithms for low-resource natural language understanding tasks, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Hong Kong, China, 2019, pp. 1192-1197.
[11] J. Zhou, Y. Lu, H. Dai, H. Wang, H. Xiao, Sentiment analysis of Chinese microblog based on stacked bidirectional LSTM, *IEEE Access*, Vol. 7, pp. 38856-38866, March, 2019.
[12] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jone, A. N. Gomez, Ł. Kaiser, I. Polosukhin, *Attention is all you need*, Advances in neural information processing systems 30, Long Beach, CA, USA, 2017, pp. 5998-6008.
[13] S. Lai, L. Xu, K. Liu, J. Zhao, Recurrent convolutional neural networks for text classification, *Twenty-ninth AAAI conference on artificial intelligence(AAAI)*, Austin, Texas, USA, 2015, pp. 2267-2273.
[14] Y. Zhang, B. Wallace, A Sensitivity Analysis of (and Practitioners' Guide to) Convolutional Neural Networks for Sentence Classification, *Proceedings of the Eighth International Joint Conference on Natural Language Processing (IJCNLP)*, Taipei, Taiwan, 2017, pp. 253-263.

Biographies

Azragul Yusup was born in 1987 in Urumqi, Xinjiang, China. in 2016, she studied for her PhD in Computer Application Technology at the University of Chinese Academy of Sciences. Currently, she is working as an associate professor in Xinjiang Normal University, School of Computer Science and Technology, with research interests in natural language processing and computational linguistics.



Degang Chen was born in 1996 in Guiyang, Guizhou, China. His research interests are in text sentiment analysis and image recognition.

Yifei Ge was born in 1998 in Suqian, Jiangsu. His research interests are in natural language processing named entity recognition.

Hongliang Mao was born in 1995 in Dingxi, Gansu. His research interests are in natural language processing named entity recognition.

Nujian Wang was born in 1995 in Qingdao, Shandong Province, China. His research interests are in textual sentiment analysis.